AK

PCT

WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 : G06T 15/70

A1

(11) International Publication Number: WO 99/46734

(43) International Publication Date: 16 September 1999 (16.09.99)

(21) International Application Number: PCT/US99/05289

(22) International Filing Date: 11 March 1999 (11.03.99)

(30) Priority Data:
60/077,565 11 March 1998 (11.03.98) US

(71) Applicant: ENTROPIC, INC. [US/US]; Suite G-100, 400 North Capitol Street, N.W., Washington, DC 20001 (US).

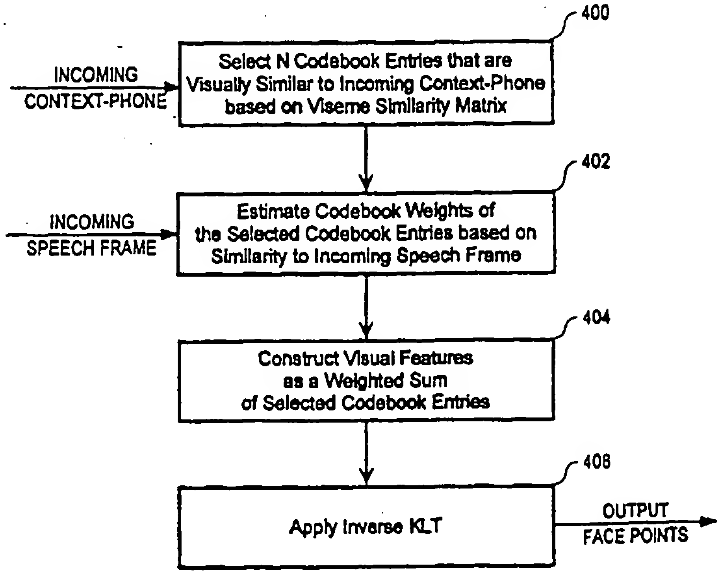
(72) Inventors: ARSLAN, Levent; Vogazici Universitesi, Bebek, 80815 Istanbul (TR). TALKIN, David; 420 N. Mary Avenue, Sunnyvale, CA 94086 (US).

(74) Agents: CARLSON, Stephen, C. et al.; McDermott, Will & Emery, 600 13th Street, N.W., Washington, DC 20005-3096 (US).

(81) Designated States: AU, CA, IL, JP, European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).

Published
With international search report.
Before the expiration of the time limit for amending the claims and to be republished in the event of the receipt of amendments.

(54) Title: FACE SYNTHESIS SYSTEM AND METHODOLOGY


INCOMING CONTEXT-PHONE
Select N Codebook Entries that are Visually Similar to Incoming Context-Phone based on Viseme Similarity Matrix
400
INCOMING SPEECH FRAME
Estimate Codebook Weights of the Selected Codebook Entries based on Similarity to Incoming Speech Frame
402
Construct Visual Features as a Weighted Sum of Selected Codebook Entries
404
Apply Inverse KLT
408
OUTPUT FACE POINTS

(57) Abstract

A system and method for synthesizing a facial image, compares a speech frame from an incoming speech signal with acoustic features stored within visually similar entries in an audio–visual codebook to produce a set of weights. The audio–visual codebook also stores visual features corresponding to the acoustic features. A composite visual feature is generated as a weighted sum of the corresponding visual features, from which the facial image is synthesized. The audio–visual codebook may include multiple samples of the acoustic and visual features for each entry, which corresponds to a sequence of one or more phonemes.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	ES	Spain	LS	Lesotho	SI	Slovenia
AM	Armenia	FI	Finland	LT	Lithuania	SK	Slovakia
AT	Austria	FR	France	LU	Luxembourg	SN	Senegal
AU	Australia	GA	Gabon	LV	Latvia	SZ	Swaziland
AZ	Azerbaijan	GB	United Kingdom	MC	Monaco	TD	Chad
BA	Bosnia and Herzegovina	GE	Georgia	MD	Republic of Moldova	TG	Togo
BB	Barbados	GII	Ghana	MG	Madagascar	TJ	Tajikistan
BE	Belgium	GN	Guinea	MK	The former Yugoslav	TM	Turkmenistan
BF	Burkina Faso	GR	Grecce		Republic of Macedonia	TR	Turkey
BG	Bulgaria	HU	Hungary	ML	Mali	TT	Trinidad and Tobago
BJ	Benin	IE	Ireland	MN	Mongolia	UA	Ukraine
BR	Brazil	IL	Israel	MR	Mauritania	UG	Uganda
BY	Belarus	IS	Iceland	MW	Malawi	US	United States of America
CA	Canada	IT	Italy	MX	Mexico	UZ	Uzbekistan
CF	Central African Republic	JP	Japan	NE	Niger	VN	Viet Nam
CG	Congo	KE	Kenya	NL	Netherlands	YU	Yugoslavia
CII	Switzerland	KG	Kyrgyzstan	NO	Norway	ZW	Zimbabwe
CI	Côte d'Ivoire	KP	Democratic People's	NZ	New Zealand		
CM	Cameroon		Republic of Korea	PL	Poland		
CN	China	KR	Republic of Korea	PT	Portugal		
CU	Cuba	KZ	Kazakstan	RO	Romania		
CZ	Czech Republic	LC	Saint Lucia	RU	Russian Federation		
DE	Germany	LI	Liechtenstein	SD	Sudan		
DK	Denmark	LK	Sri Lanka	SE	Sweden		
EE	Estonia	LR	Liberia	SG	Singapore		

FACE SYNTHESIS SYSTEM AND METHODOLOGY

RELATED APPLICATIONS

This application claims the benefit of U.S. Provisional Application No. 60/077,565 entitled "Face Synthesis System and Methodology" filed on March 11, 1998 by Levent M. Arslan and David Talkin, the contents of which are incorporated by reference herein.

This application contains subject matter related to the commonly assigned, co-pending PCT patent application No. PCT/US98/01538 entitled "Voice Conversion System and Methodology" filed on January 27, 1998 naming Levent M. Arslan and David Talkin as inventors, the contents of which are incorporated by reference herein.

FIELD OF THE INVENTION

The present invention relates to audiovisual systems and, more particularly, to a system and methodology for face synthesis.

BACKGROUND OF THE INVENTION

Recently there has been significant interest in face synthesis. Face synthesis refers to the generation of a facial image in accordance with a speech signal, so that it appears to a viewer that the facial image is speaking the words uttered in the speech signal. There are many applications of face synthesis including film dubbing, cartoon character animation, interactive agents, and multimedia entertainment.

Face synthesis generally involves a database of facial images in correspondence with distinct sounds of a language. Each distinct sound of the language is referred to as a "phoneme," and during pronunciation of a phoneme, the mouth and lips of a face form a characteristic, visible configuration, referred to as a "viseme." Typically, the facial image database includes a "codebook" that maps each phoneme of a language to a corresponding viseme. Accordingly, the input speech text is segmented into phonemes, and the corresponding viseme for each phoneme is sequentially fetched from the database and displayed.

Realistic image quality is an important concern in face synthesis, and transitions from one sound to the next are particularly difficult to implement in a life-like manner because the mouth and lips are moving during the course of pronouncing a sound. In one approach, the mathematical routines are employed to interpolate a series of intermediate images from one viseme at one phoneme to the next. Such an approach, however, can result in an unnatural or distorted appearance, because the movements from one mouth and lip configuration to another are often non-linear.

In general, it is practical to store only a restricted number of phoneme/viseme sequences in the codebook. For example, image quality may be improved by storing visemes for all the allophones of a phoneme. An allophone of a phoneme is a slight, non-contrastive variation in pronunciation of the phoneme. A similar issue occurs in applying a face synthesis system originally developed for one language to speech in another language, because the other language includes additional phonemes lacking in the original language. Furthermore, the precise shape of a viseme is often dependent on the neighboring visemes, and there has been some interest in using sequences of phonemes of a given length, such as diphones.

Augmenting the codebook for every possible allophone, foreign phoneme, and phoneme sequences with their corresponding visemes consumes an unacceptably large amount of storage. In a common approach, aliasing techniques are employed in which visemes for a missing phoneme or sequence of phoneme are replaced by existing visemes in the codebook. Aliasing, however, tends to introduce artifacts at the frame boundaries, thereby reducing the realism of the final image.

SUMMARY OF THE INVENTION

Accordingly, there exists a need for a face synthesis system and methodology that generates realistic facial images. In particular, there is a need for handling transitions from one viseme to the next with improved realism. Furthermore, a need exists for generating realistic facial images for sequences of phonemes that are missing the codebook or for foreign language phonemes.

These and other needs are addressed by a method and computer-readable medium bearing instructions for synthesizing a facial image, in which a speech frame from an incoming speech signal is compared against acoustic features stored within an audio-visual codebook to produce a set of weights. These weights are used to generate a composite visual feature based on visual features corresponding to the acoustic features, and the composite visual feature is then used to synthesize a facial image. Generating a facial image based on a weighted composition of other images is a flexible approach that allows for more realistic facial images.

For example, more realistic viseme transitions during the course of pronunciation may be realized by using multiple samples of the acoustic and visual features for each entry in the audio-visual codebook, taken during the course of pronouncing a sound. Visemes for foreign phonemes can be generated by combining visemes from a combination of audio-visual codebook entries that correspond to native phonemes. For context-sensitive audio-visual codebooks with a restricted number of phoneme sequences, a weighted combination of features from visually similar phoneme sequences allows for a realistic facial image to be produced for a missing phoneme sequence.

In one embodiment, both the aforementioned aspects are combined so that each entry in the audio-visual codebook corresponds to a phoneme sequence and includes multiple samples of acoustic and visual features. In some embodiments, the acoustic features may be implemented by a set of line spectral frequencies and the visual features by the principal components of a Karhunen-Loewe transform of face points.

Additional objects, advantages, and novel features of the present invention will be set forth in part in the description that follows, and in part, will become apparent upon examination or may be learned by practice of the invention. The objects and advantages of the invention may be realized and obtained by means of the instrumentalities and combinations particularly pointed out in the appended claims.

BRIEF DESCRIPTION OF THE DRAWINGS

The present invention is illustrated by way of example, and not by way of limitation, in the figures of the accompanying drawings and in which like reference numerals refer to similar elements and in which:

FIG. 1 schematically depicts a computer system that can implement the present invention;

FIGS. 2(a) and 2(b) depict the influence of a modification to the first and second principal components, respectively, of face point data.

FIG. 3 depicts a viseme similarity matrix 300 corresponding to phoneme in American English.

FIG. 4 is a flowchart illustrating a face synthesis process in accordance with one embodiment of the present invention.

DESCRIPTION OF THE PREFERRED EMBODIMENT

A method and system for face synthesis are described. In the following description, for the purposes of explanation, numerous specific details are set forth in order to provide a thorough understanding of the present invention. It will be apparent, however, to one skilled in the art that the present invention may be practiced without these specific details. In other instances, well-known structures and devices are shown in block diagram form in order to avoid unnecessarily obscuring the present invention.

HARDWARE OVERVIEW

Figure 1 is a block diagram that illustrates a computer system 100 upon which an embodiment of the invention may be implemented. Computer system 100 includes a bus 102 or other communication mechanism for communicating information, and a processor (or a plurality of central processing units working in cooperation) 104 coupled with bus 102 for processing information. Computer system 100 also includes a main memory 106, such as a random access memory (RAM) or other dynamic storage device, coupled to bus 102 for storing information and instructions to be executed by processor 104. Main memory 106 also may be used for storing temporary variables or other intermediate information during execution of instructions to be executed by processor 104. Computer system 100 further includes a read only memory (ROM) 108 or other static storage device coupled to bus 102 for storing static information and instructions for processor 104. A storage device 110, such as a magnetic disk or optical disk, is provided and coupled to bus 102 for storing information and instructions.

Computer system 100 may be coupled via bus 102 to a display 111, such as a cathode ray tube (CRT), for displaying information to a computer user. An input device 113, including alphanumeric and other keys, is coupled to bus 102 for communicating information and command selections to processor 104. Another type of user input device is cursor control 115, such as a mouse, a trackball, or cursor direction keys for communicating direction information and command selections to processor 104 and for controlling cursor movement on display 111. This input device typically has two degrees of freedom in two axes, a first axis (e.g., *x*) and a second axis (e.g., *y*), that allows the device to specify positions in a plane. For audio output and input, computer system 100 may be coupled to a speaker 117 and a microphone 119, respectively.

The invention is related to the use of computer system 100 for face synthesis. According to one embodiment of the invention, face synthesis is provided by computer system 100 in response to processor 104 executing one or more sequences of one or more instructions contained in main memory 106. Such instructions may be read into main memory 106 from another computer-readable medium, such as storage device 110. Execution of the sequences of instructions contained in main memory 106 causes processor 104 to perform the process steps described herein. One or more processors in a multi-processing arrangement may also be employed to execute the sequences of instructions contained in main memory 106. In alternative embodiments, hard-wired circuitry may be used in place of or in combination with software instructions to implement the invention. Thus, embodiments of the invention are not limited to any specific combination of hardware circuitry and software.

The term "computer-readable medium" as used herein refers to any medium that participates in providing instructions to processor 104 for execution. Such a medium may take many forms, including but not limited to, non-volatile media, volatile media, and transmission media. Non-volatile media include, for example, optical or magnetic disks, such as storage device 110. Volatile media include dynamic memory, such as main memory 106. Transmission media include coaxial cables, copper wire and fiber optics, including the wires that comprise bus 102. Transmission media can also take the form of acoustic or light waves, such as those generated during radio frequency (RF) and infrared (IR) data communications. Common forms of computer-readable media include, for example, a floppy disk, a flexible disk, hard disk, magnetic tape, any other magnetic medium, a CD-ROM, DVD, any other optical medium, punch cards, paper tape, any other physical medium with patterns of holes, a RAM, a PROM, and EPROM, a FLASH-EPROM, any other memory chip or cartridge, a carrier wave as described hereinafter, or any other medium from which a computer can read.

Various forms of computer readable media may be involved in carrying one or more sequences of one or more instructions to processor 104 for execution. For example, the instructions may initially be borne on a magnetic disk of a remote computer. The remote computer can load the instructions into its dynamic memory and send the instructions over a telephone line using a modem. A modem local to computer system 100 can receive the data on the telephone line and use an infrared transmitter to convert the data to an infrared signal. An infrared detector coupled to bus 102 can receive the data carried in the infrared signal and place the data on bus 102. Bus 102 carries the data to main memory 106, from which processor 104 retrieves and executes the instructions. The instructions received by main memory 106 may optionally be stored on storage device 110 either before or after execution by processor 104.

Computer system 100 also includes a communication interface 120 coupled to bus 102. Communication interface 120 provides a two-way data communication coupling to a network link 121 that is connected to a local network 122. Examples of communication interface 120 include an integrated services digital network (ISDN) card, a modem to provide a data communication connection to a corresponding type of telephone line, and a local area network (LAN) card to provide a data communication connection to a compatible LAN. Wireless links may also be implemented. In any such implementation, communication interface 120 sends and receives electrical, electromagnetic or optical signals that carry digital data streams representing various types of information.

Network link 121 typically provides data communication through one or more networks to other data devices. For example, network link 121 may provide a connection through local network 122 to a host computer 124 or to data equipment operated by an Internet Service Provider (ISP) 126. ISP 126 in turn provides data communication services through the world wide packet data communication network, now commonly referred to as the “Internet” 128. Local network 122 and Internet 128 both use electrical, electromagnetic or optical signals that carry digital data streams. The signals through the various networks and the signals on network link 121 and through communication interface 120, which carry the digital data to and from computer system 100, are exemplary forms of carrier waves transporting the information.

Computer system 100 can send messages and receive data, including program code, through the network(s), network link 121, and communication interface 120. In the Internet example, a server 130 might transmit a requested code for an application program through Internet 128, ISP 126, local network 122 and communication interface 118. In accordance with the invention, one such downloaded application provides for face synthesis as described herein. The received code may be executed by processor 104 as it is received, and/or stored in storage device 110, or other non-volatile storage for later execution. In this manner, computer system 100 may obtain application code in the form of a carrier wave.

AUDIO-VISUAL CODEBOOK

In accordance with an embodiment of the present invention, an off-line training phase is undertaken as a preliminary step to generate an audio-visual codebook and, preferably, a viseme similarity matrix 300. An audio-visual codebook is data structure that contains entries corresponding to a single phoneme or to a central phoneme in sequence of phonemes, called a "context phone." Each entry includes one or more acoustic features for the phoneme and the corresponding visual features of a related viseme.

The off-line training phase involves collecting data from a test subject by recording the synchronized speech and face point trajectories of the subject. According to one training approach, the subject is asked to utter words, phrases, and sentences for which an orthographic transcription is prepared. The recorded acoustic and visual data are then processed and stored in entries in the audio-visual codebook. The number of entries in the audio-visual codebooks will vary from implementation to implementation and generally depends on a desired trade-off between face synthesis quality and computational performance.

In one embodiment, the acoustic data is sampled at an appropriate frequency such as 16 kHz and automatically segmented using, for example, a forced alignment to a phonetic translation of the orthographic transcription within an HMM framework using Mel-cepstrum coefficients and delta coefficients as described in more detail in C. Wightman & D. Talkin, *The Aligner User's Manual*, Entropic Reseach Laboratory, Inc., Washington, D.C., 1994. Preferably, the sampled voice data is converted into line spectral frequencies, which can be estimated quite reliably and have a fixed range useful for real-time digital signal processing. The line spectral frequency values for the audio-visual codebook can be obtained by first determining the linear predictive coefficients $\mathbf{a}_k$ for the sampled signal according to well-known techniques in the art. For example, specialized hardware, software executing on a general purpose computer or microprocessor, or a combination thereof, can ascertain the linear predictive coefficients by such techniques as square-root or Cholesky decomposition, Levinson-Durbin recursion, and lattice analysis introduced by Itakura and Saito.

In one embodiment, the visual data is obtained as 52 "face points" in three-dimensional space corresponding to points on the subject's face. Since each face point represents *x*, *y*, and *z* coordinates, the total number of face point parameters is 156, thereby constituting a 156-dimensional face point vector. An appropriate transformation technique, such as the Karhunen-Loewe transform, is applied to the face point vector to obtain its principal components. Since the points of a face are highly correlated, a significant reduction in dimensionality can be achieved with only minor distortion. A useful property of using principal components to represent visual features is that the principal components designate the directions that correspond to the most correlated movements. Therefore, modifying the weights of the principal components can be used to animate the underlying face with realistic motions.

For example, the eigenvector with the largest eigenvalue was found to correspond with the movement of the lower jaw, which involves the largest set of correlated points in a speaker's face. Thus, modifying just the first principal component results in moving the lower lip and jaw trajectories. The second principal component was found to correspond with the movement of the sides of the mouth. FIGS. 2(a) and 2(b) depict the effect of adjusting only the first and second principal components of a face point vector, respectively, wherein dark curves represent an original face point trajectory and light curves represent the adjusted face point trajectory.

In accordance with one aspect, each phoneme segmented from the speech data is tagged with a "context-phone" symbol indicating the context of phoneme. Specifically, the context-phone symbol indicates the phoneme in the center and one or more neighboring phonemes on either side of the center phoneme in the speech data. For example, the phoneme /eh/ in the word "whenever" has a context-phone symbol of /w_eh_n_eh_v_axr_f/ that includes the three closest neighbors on either side. (The rightmost /f/ phoneme belongs to the following word that begins with an 'f' or 'ph'.) Use of context phones, which form a sequence of phonemes including a center phoneme and neighboring phonemes, allows appropriate context-specific visemes to be generated.

In accordance with another aspect, each phoneme in the training data is labeled with multiple, uniformly spaced time locations, for example at five locations, within the course of articulation of the phoneme. The acoustic and visual features, *e.g.* line spectral frequencies and Karhunen-Loewe principal components, are stored in the audio-visual codebook entry for the phoneme or context-phone. Use of multiple acoustic and visual features allows for a smooth and realistic sequence of visemes to be generated during the course of phoneme articulation.

Thus, the audio-visual codebook includes a number of entries corresponding to a phoneme or a center phoneme and including one or more acoustic features and one or more corresponding visual features. The audio-visual codebook can be used to generate facial images by comparing an incoming speech frame with the acoustic features in the entries to estimate weights for each of the compared acoustic features. The corresponding visual features are combined as a weighted sum to produce a composite visual feature, which is converted into a facial image. Although performing this process for all the entries in the audio-visual codebook results in a very high quality output, it is desirable improve the performance of this process.

VISEME SIMILARITY MATRIX

In one embodiment, the performance can be significantly improved if phonetic information is known a priori about the incoming speech data being synthesized into facial images. Specifically, several entries in the audio-visual codebook are selected whose phonemes or context-phones are most visually similar to the phoneme being pronounced in each incoming speech frame. Thus, the total number of entries that are compared with the acoustic feature of the incoming speech frame is reduced to only a few of the most visually similar entries. This selection reduces the computational overhead of the system and improves the overall performance of the face synthesis process.

Since, in practice, the training data will not include all possible context-phones of a given length (or all foreign phonemes and allophones), it is desirable to have some method of associating an unseen context-phone with visually similar entries in the audio-visual codebook. One visually similar measure is based on the Euclidean distance of the principal components of face data. This similarity measure can be automatically generated from the training data and stored in a viseme similarity matrix 300 by estimating an average principal component vector $\mathbf{m}_k$ for each phoneme from the various instances of the phoneme in the training data, as follows:

$$\mathbf{m}_k = \frac{1}{T}\sum_{t=1}^{T}\mathbf{p}_{kt}\,, \qquad k \in 1..K, \qquad (1)$$

where K represents the total number of phoneme in the language, T represent the total number of the kth phonemes in the training data, and p_{kl} represents the lth principal component vector that is associated with the kth phoneme. Given the average principal component vectors $\mathbf{m}_k$, the Euclidean distance between each pair of phonemes is calculated as:

$$d_{ik} = \|\mathbf{m}_i - \mathbf{m}_k\|, \qquad i, k \in 1..K, \tag{2}$$

Based on the calculated Euclidean distances, the viseme similarity measure s_{ik} is derived as follows:

$$s_{ik} = e^{-50 d_{ik}}, \qquad i, k \in 1..K, \tag{3}$$

One property of this formulation is that viseme similarity values s_{ik} will range between 0 and 1. FIG. 3 depicts a gray scale image of one viseme similarity matrix 300 corresponding to phonemes of American English, wherein darker points represent a higher level of visual similarity. For example, the most visually similar phoneme to /b/ was identified to be /p/. In general, it has been found that entries in the viseme similarity matrix 300 agree with intuitive expectations.

The viseme similarity matrix 300 can be used directly to determine the visual similarity of two phonemes, but a more involved procedure is used to estimate a visual similarity measure between two context-phones representing a sequence of phonemes. Preferably, the center phoneme should have the highest influences with decreasing influence for the phonemes that are more remote from the center phoneme. One procedure to estimate the visual similarity of context-phones can be formulated as follows:

$$v_j = s_{cj} \sum_{i=1}^{C} 10^{-i}(s_{lij} + s_{rij}), \quad j \in 1..L, \tag{4}$$

where C is the level context information (*i.e.* the number of neighboring phonemes on each side), L is the total number of content-phones in the audio-visual codebook, s_{lij} is the similarity between the ith left phoneme of the subject context-phone and the jth context-phone in the audio-visual codebook, and s_{rij} is the similarity between the ith right phoneme of the subject context-phone and the jth context-phone in the audio-codebook was configured to store entries for only a single phone, then the viseme similarity matrix 300 is consulted directly to obtain the visual similarity measure.

Based on the determined visual similarity measures, the N most visually similar entries of the audio-visual codebook are selected. The best value for N will vary from implementation to implementation, depending on factors such as the length of the phoneme sequence of the context-phones and the desired performance/realism tradeoff for a given set of training data. Generally, however, the value of N ranges from about four to about sixteen, and may in fact be a user-configurable parameter.

At step 402, the incoming speech frame is converted into acoustic features suitable for comparison with the acoustic features stored in the audio-visual codebook. For example, the incoming speech frame may be converted into line spectral frequencies and compared with a line spectral frequencies set stored in the audio-visual codebook. In some embodiment, a plurality of samples, such as five samples, are stored for each entry in the audio-visual codebook. The result of the acoustic feature comparison is a weight, wherein a higher weight is assigned for more acoustically similar samples. A variety of techniques for producing the weight based on the comparison may be employed but the present invention is not limited to any particular weight.

One weighting technique is described in the commonly assigned, co-pending PCT patent application No. PCT/US98/01538 entitled "Voice Conversion System and Methodology." As described therein, codebook weights $\mathbf{v}_i$ are estimated by comparing the input line spectral frequency vector $\mathbf{w}_k$ with each acoustic feature sample, $\mathbf{S}_i$ in the audio-visual codebook to calculate a corresponding distance d_i:

$$d_i = \sum_{k=1}^{P} \mathbf{h}_k \, | \mathbf{w}_k - \mathbf{S}_{ik} |, i \in 1..L \qquad (5)$$

where L is the codebook size. The distance calculation may include a weight factor $\mathbf{h}_k$, which is based on a perceptual criterion wherein closely spaced line spectral frequency pairs, which are likely to correspond to formant locations, are assigned higher weights:

$$\mathbf{h}_k = \frac{e^{-0.05|K-k|}}{\min(|\mathbf{w}_k - \mathbf{w}_{k-1}|, |\mathbf{w}_k - \mathbf{w}_{k+1}|)}, k \in 1..P \tag{6}$$

where K is 3 for voiced sounds and 6 for unvoiced, since the average energy decreases (for voiced sounds) and increases (for unvoiced sounds) with increasing frequency. Based on the calculated distances d_i, the normalized codebook weights $\mathbf{v}_i$ are obtained as follows:

$$\mathbf{v}_i = \frac{e^{-\gamma d_i}}{\sum_{l=1}^{L} e^{-\gamma d_l}}, i \in 1..L \tag{7}$$

where the value of γ for each frame is found by an incremental search in the range of 0.2 to 2.0 with the criterion of minimizing the perceptual weighted distance between the approximated line spectral frequency vector $\mathbf{v}S_k$ and the input line spectral frequency vector $\mathbf{w}_k$. These weights may be further adjusted as also described in pending PCT patent application No. PCT/US98/01538.

At step 404, a composite visual feature is constructed from the weights and the corresponding visual features of the selected audio-visual codebook entries, for example, as a weighted sum or linear combination of the principal components of the facial data samples. For example, the composite visual feature may be calculated as follows:

$$\tilde{\mathbf{p}} = \sum_{n=1}^{5N} \mathbf{v}_n \mathbf{p}_n . \tag{8}$$

In one embodiment, a plurality of visual features is stored for each entry in the audio-visual codebook at different points in time during the articulation of the sound corresponding to the entry. Thus, the weighted sum will include all the visual samples for the audio-visual codebook entry, thereby producing facial data that more realistically tracks the movement of the mouth and lips during speaking.

At step 406, the composite visual feature is converted into the desired facial data. For example, if principal components obtained from a Karhunen-Loewe transformation are used to represent the visual features, then an inverse Karhunen-Loewe transformation is applied on the composite principal components to produce face points as output. These face points can be converted to the facial image by known techniques.

Accordingly, a face synthesis system and methodology is described wherein realistic facial images are produced in accordance with an input speech signal. Specifically, a composite visual feature is generated from entries in an audio-visual codebook according to weights identified by comparing the incoming acoustic features with the audio-visual codebook acoustic features. Consequently, realistic output is attained for viseme transitions, for highly context-dependent situations, and even for foreign language phonemes without requiring the audio-visual codebook to store an enormous amount of training samples.

While this invention has been described in connection with what is presently considered to be the most practical and preferred embodiment, it is to be understood that the invention is not limited to the disclosed embodiment, but on the contrary, is intended to cover various modifications and equivalent arrangements included within the spirit and scope of the appended claims.

CLAIMS

WHAT IS CLAIMED IS:

1. A method of synthesizing a facial image in accordance with a speech signal, comprising the steps of:

comparing a speech frame of the speech signal with a plurality of acoustic features within an audio-visual codebook to produce therefrom a plurality of weights;

generating a composite visual feature based on the weights and a plurality of visual features corresponding to the acoustic features; and

synthesizing the facial image based on the composite visual feature.

2. The method of claim 1, wherein:

the audio-visual codebook contains entries each including a plurality of acoustic features and a plurality of corresponding visual features; and

comparing the speech frame includes comparing the speech frame with the acoustic features from an entry within the audio-visual codebook.

3. The method of claim 1, wherein:

the audio-visual codebook contains entries each including an acoustic feature and a corresponding visual feature; and

comparing the speech frame includes comparing the speech frame with acoustic features from a plurality of selected entries within the audio-visual codebook.

4. The method of claim 3, wherein the speech signal includes a sequence of speech frames correlated with a sequence of phonemes, said method further comprising:

determining a visual similarity measure between a phoneme in the sequence that is correlated to the speech frame and the entries in the audio-visual codebook, said entries in the audio-visual codebook corresponding to a phoneme; and

selecting the selected entries from the audio-visual codebook based on the visual similarity measure; and

5. The method of claim 3, wherein the speech signal includes a sequence of speech frames correlated with a sequence of phonemes, said method further comprising:

determining visual similarity measures between a phoneme in the sequence with neighboring phonemes thereof and the entries in the audio-visual codebook, said entries in the audio-visual codebook corresponding to a series of phonemes having a center phoneme with neighboring phonemes thereof; and

selecting the selected entries in the audio-visual codebook based on the determined visual similarities.

6. The method of claim 5, wherein determining the visual similarity measures includes calculating Euclidean distances between each of sets of principal components of facial data corresponding to the phoneme in the sequence with the neighboring phonemes thereof and principle component samples of facial data corresponding to the center phoneme with the neighboring phonemes thereof.

7. The method of claim 5, wherein determining the visual similarity measures includes accessing a visual similarity matrix containing elements based on Euclidean distances between each of sets of principal components of facial data corresponding to the phoneme in the sequence with the neighboring phonemes thereof and principle component samples of facial data corresponding to the center phoneme with the neighboring phonemes thereof.

8. The method of claim 1, wherein the acoustic feature includes a line spectral frequencies set and the visual feature includes a set of principal components of facial data derived from face point samples.

9. A method of synthesizing a facial image in accordance with a speech signal, said speech signal including a sequence of speech frames correlated with a sequence of phonemes, said method comprising the steps of:

determining visual similarity measures between a phoneme in the sequence with neighboring phonemes thereof and entries in an audio-visual codebook, said entries in the audio-visual codebook corresponding to a series of phonemes having a center phoneme with neighboring phonemes thereof and including a plurality of acoustic features and a plurality of corresponding visual features;

selecting a plurality of the entries in the audio-visual codebook based on the determined visual similarities;

comparing a speech frame of the speech signal with the acoustic features from entries to produce therefrom a plurality of weights;

generating a composite visual feature based on the visual features of the entries and the weights; and

synthesizing the facial image based on the composite visual feature.

10. A computer-readable medium bearing instructions for synthesizing a facial image in accordance with a speech signal, said instructions arranged, when executed by one or more processors, to cause the one or more processors to perform the steps of:

comparing a speech frame of the speech signal with a plurality of acoustic features within an audio-visual codebook to produce therefrom a plurality of weights;

generating a composite visual feature based on the weights and a plurality of visual features corresponding to the acoustic features; and

synthesizing the facial image based on the composite visual feature.

11. A computer-readable medium bearing instructions for synthesizing a facial image in accordance with a speech signal, said speech signal including a sequence of speech frames correlated with a sequence of phonemes, said instructions arranged, when executed by one or more processors, to cause the one or more processors to perform the steps of:

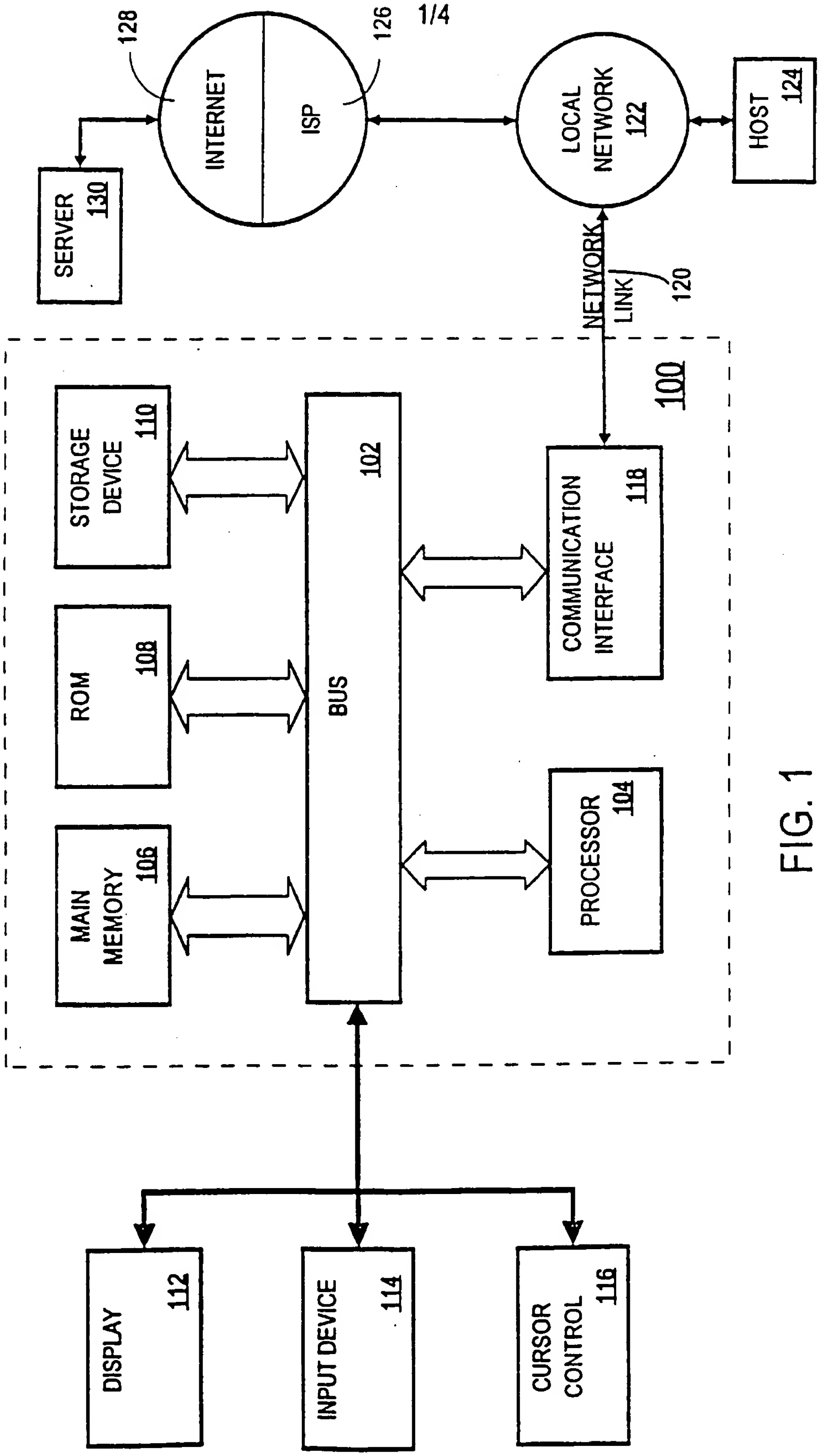
determining visual similarity measures between a phoneme in the sequence with neighboring phonemes thereof and entries in an audio-visual codebook, said entries in the audio-visual codebook corresponding to a series of phonemes having a center phoneme with neighboring phonemes thereof and including a plurality of acoustic features and a plurality of corresponding visual features;

selecting a plurality of the entries in the audio-visual codebook based on the determined visual similarities;

comparing a speech frame of the speech signal with the acoustic features from entries to produce therefrom a plurality of weights;

generating a composite visual feature based on the visual features of the entries and the weights; and

synthesizing the facial image based on the composite visual feature.



SERVER 130
INTERNET
128
ISP
126
LOCAL NETWORK 122
HOST 124
NETWORK LINK
120
100
STORAGE DEVICE 110
ROM 108
MAIN MEMORY 106
BUS
102
COMMUNICATION INTERFACE 118
PROCESSOR 104
DISPLAY 112
INPUT DEVICE 114
CURSOR CONTROL 116

FIG. 1

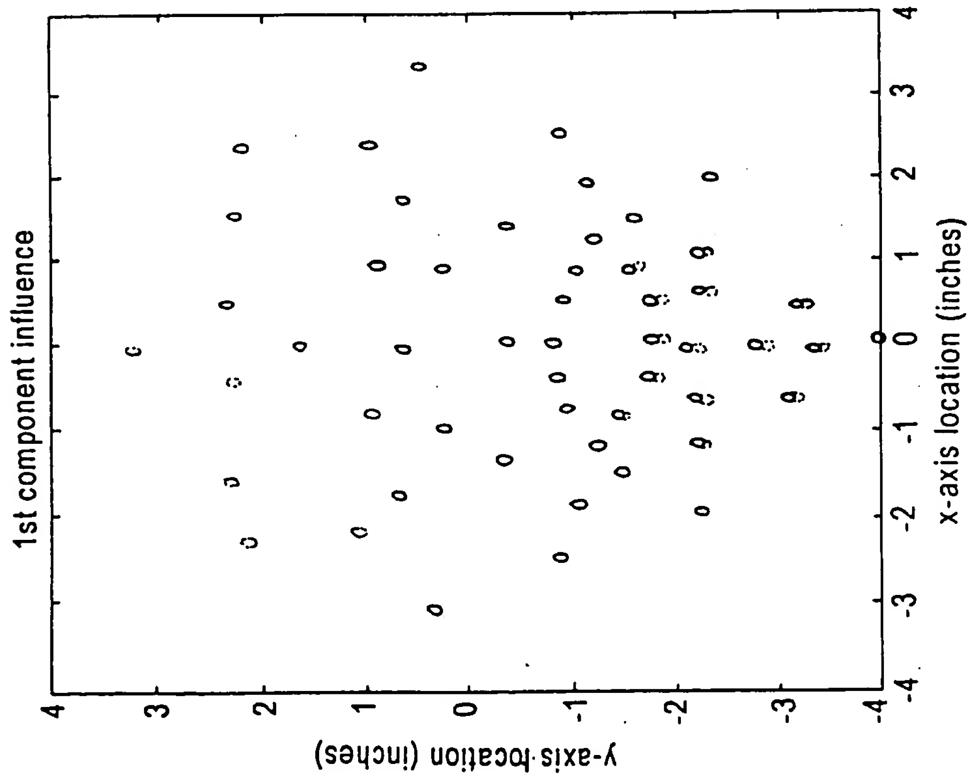

1st component influence
y-axis location (inches)
x-axis location (inches)
4
3
2
1
0
-1
-2
-3
-4
-4
-3
-2
-1
0
1
2
3
4

FIG. 2(a)

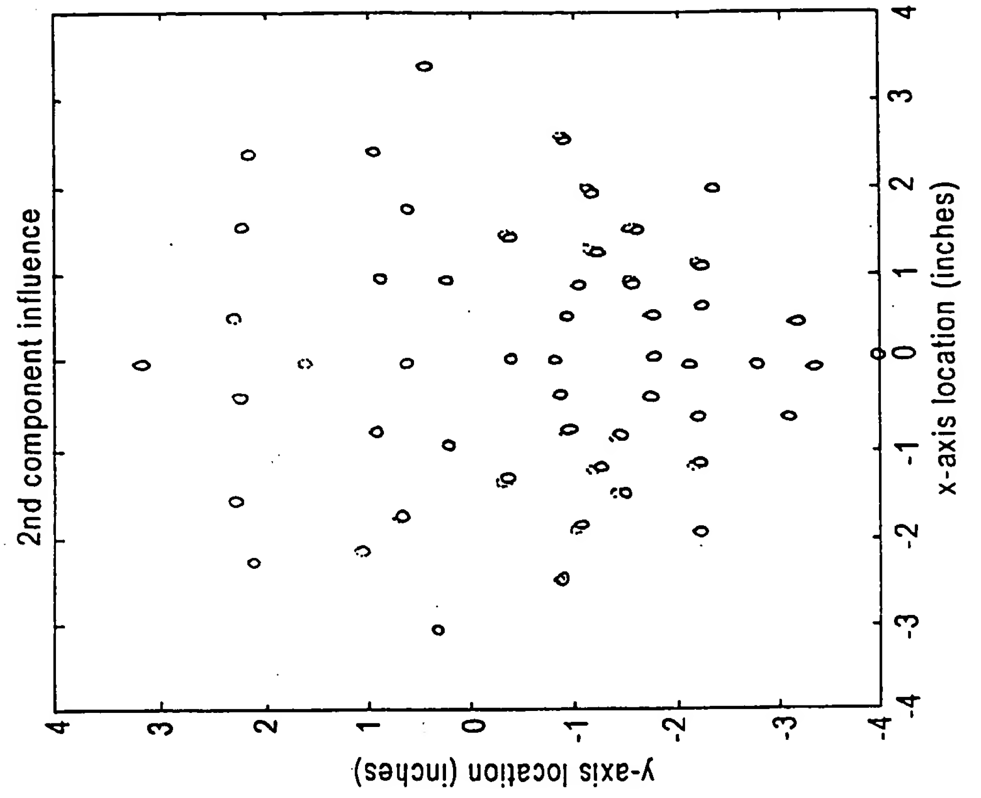

2nd component influence
y-axis location (inches)
x-axis location (inches)
4
3
2
1
0
-1
-2
-3
-4
-4
-3
-2
-1
0
1
2
3
4

FIG. 2(b)

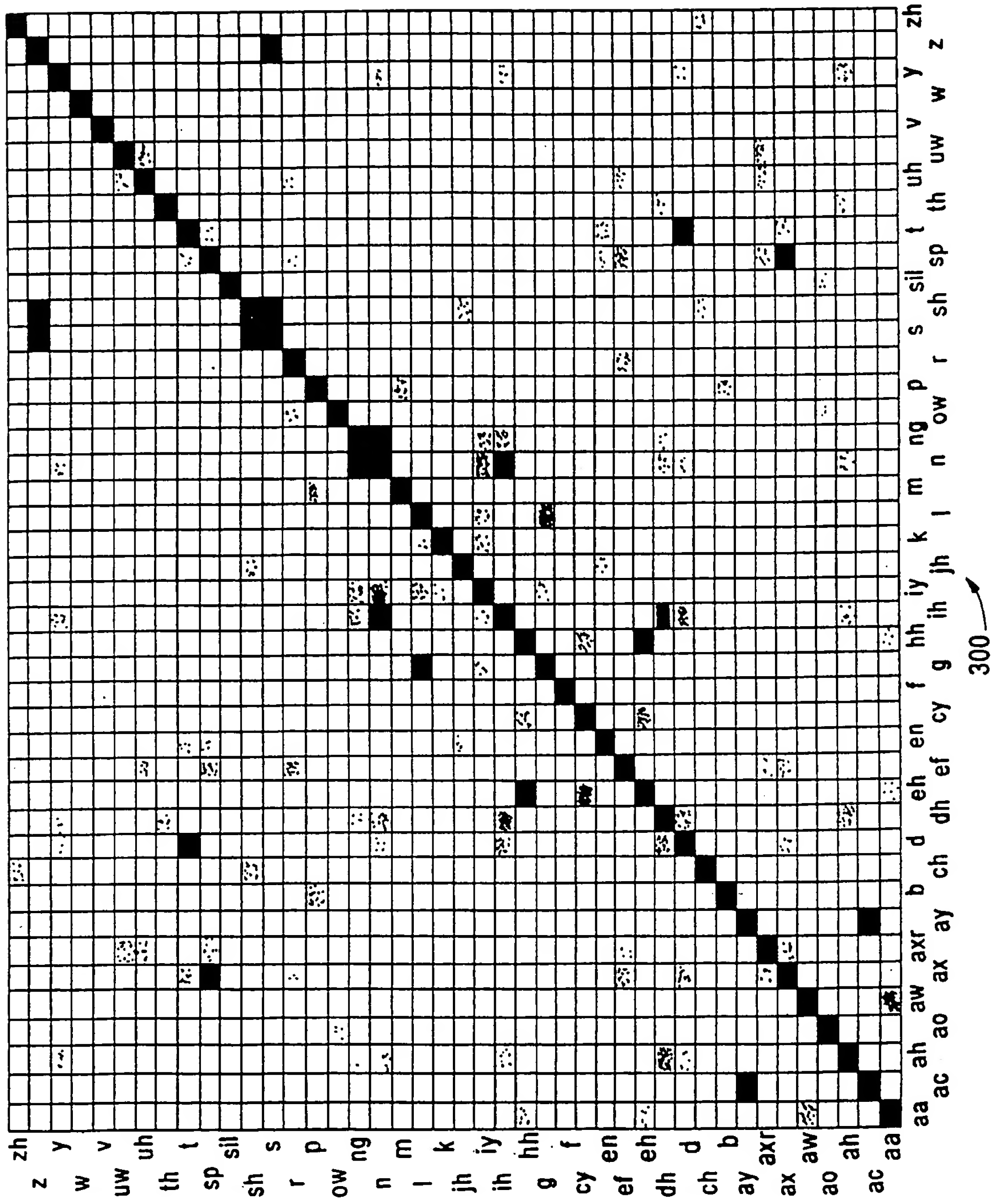

zh z y w v uw uh th t sp sil sh s r p ow ng n m l k jh iy ih hh g f cy en ef eh dh d ch b ay axr ax aw ao ah ac aa
zh z y w v uw uh th t sp sil sh s r p ow ng n m l k jh iy ih hh g f cy en ef eh dh d ch b ay axr ax aw ao ah ac aa
300

FIG. 3

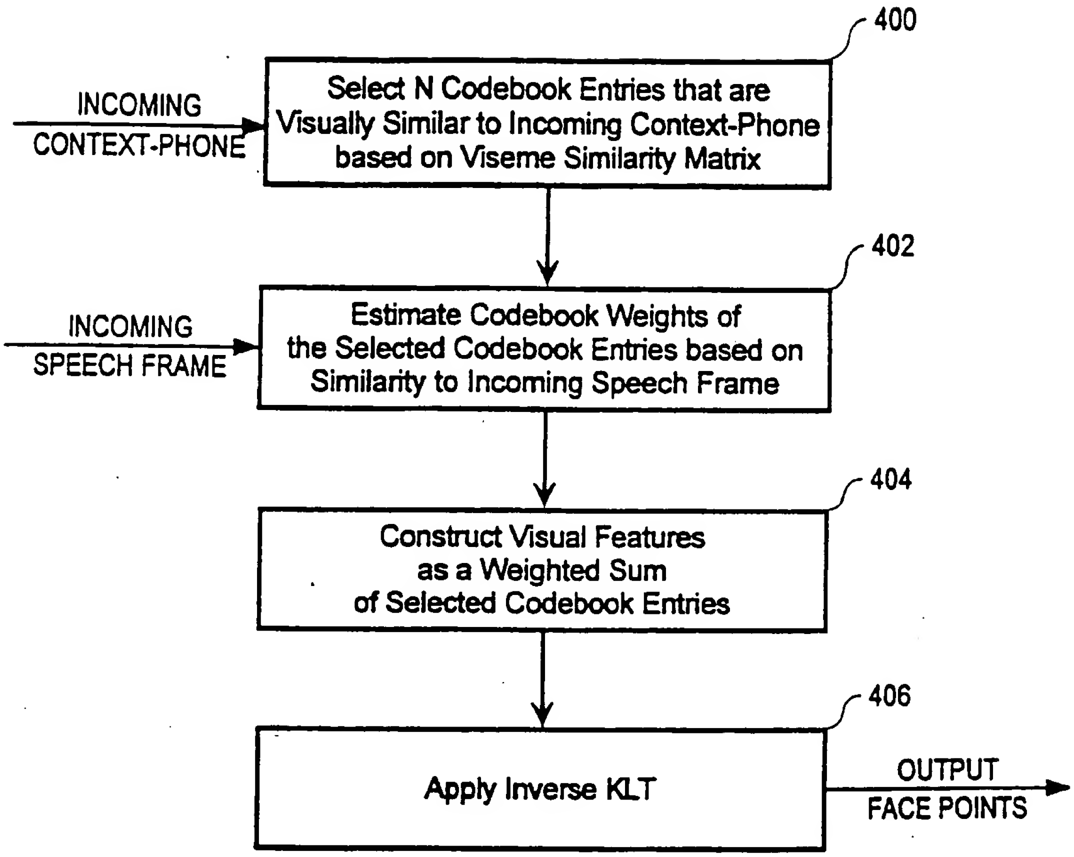

400
INCOMING
CONTEXT-PHONE
Select N Codebook Entries that are Visually Similar to Incoming Context-Phone based on Viseme Similarity Matrix
402
INCOMING
SPEECH FRAME
Estimate Codebook Weights of the Selected Codebook Entries based on Similarity to Incoming Speech Frame
404
Construct Visual Features as a Weighted Sum of Selected Codebook Entries
406
Apply Inverse KLT
OUTPUT
FACE POINTS

FIG. 4

INTERNATIONAL SEARCH REPORT

Ir ational Application No
PCT/US 99/05289

A. CLASSIFICATION OF SUBJECT MATTER
IPC 6 G06T15/70

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC 6 G06T

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	EP 0 710 929 A (AT & T CORP) 8 May 1996 (1996-05-08)	1-5,9-11
Y	abstract; claims 1,7 column 5, line 41 - line 44 ---	6-8
Y	US 4 907 276 A (ALDERSBERG SHABTAI) 6 March 1990 (1990-03-06) abstract; claims 1-3 ---	6-8
A	US 5 630 017 A (MATTHEWS III JOSEPH H ET AL) 13 May 1997 (1997-05-13) column 38, line 3 - line 17 claim 10 ---	1-11
A	GB 2 231 246 A (KOKUSAI DENSHIN DENWA CO LTD) 7 November 1990 (1990-11-07) page 4, line 19 - page 5, line 21; claims 1-3 --- -/--	1-11

☒ Further documents are listed in the continuation of box C.

☒ Patent family members are listed in annex.

° Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of the international search 21 July 1999	Date of mailing of the international search report 30/07/1999
Name and mailing address of the ISA European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax: (+31-70) 340-3016	Authorized officer Filloy Garcia, E

1

INTERNATIONAL SEARCH REPORT

Ir.   ational Application No
PCT/US 99/05289

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication,where appropriate, of the relevant passages	Relevant to claim No.
A	WO 97 36288 A (BREEN ANDREW PAUL ;BOWERS EMMA JANE (GB); BRITISH TELECOMM (GB)) 2 October 1997 (1997-10-02) page 11, paragraph 1 ---	5-7
A	SIROVICH L ET AL: "LOW-DIMENSIONAL PROCEDURE FOR THE CHARACTERIZATION OF HUMAN FACES" JOURNAL OF THE OPTICAL SOCIETY OF AMERICA - A, vol. 4, no. 3, 1 March 1987 (1987-03-01), pages 519-524, XP000522491 ISSN: 0740-3232 page 1, left-hand column, paragraph 2 ---	6-8
A	EP 0 689 362 A (AT & T CORP) 27 December 1995 (1995-12-27) column 4, line 47 - line 51 ---	1-11
A	CHOU W ET AL: "SPEECH RECOGNITION FOR IMAGE ANIMATION AND CODING" PROCEEDINGS OF THE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING (ICASSP), DETROIT, MAY 9 - 12, 1995 IMAGE AND MULTI-DIMENSIONAL SIGNAL PROCESSING/ SIGNAL PROCESSING APPLICATIONS DEVELOPMENT, vol. 4, no. CONF. 20, 9 May 1995 (1995-05-09), pages 2253-2256, XP000535403 INSTITUTE OF ELECTRICAL AND ELECTRONICS ENGINEERSISBN: 0-7803-2432-3 -----	

INTERNATIONAL SEARCH REPORT

Information on patent family members

Int tional Application No
PCT/US 99/05289

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
EP 0710929 A	08-05-1996	AU 3668095 A	16-05-1996
		CA 2162199 A	08-05-1996
		JP 8235384 A	13-09-1996
US 4907276 A	06-03-1990	NONE	
US 5630017 A	13-05-1997	US 5689618 A	18-11-1997
		US 5613056 A	18-03-1997
GB 2231246 A	07-11-1990	JP 2234285 A	17-09-1990
		JP 2518683 B	24-07-1996
WO 9736288 A	02-10-1997	AU 2167097 A	17-10-1997
		CA 2249016 A	02-10-1997
		CN 1214784 A	21-04-1999
		EP 0890168 A	13-01-1999
EP 0689362 A	27-12-1995	US 5608839 A	04-03-1997
		CA 2149068 A	22-12-1995
		JP 8023530 A	23-01-1996